Preventive Control for Power System Transient Security Based on XGBoost and DCOPF with Consideration of Model Interpretability

Songtao Zhang, Dongxia Zhang, Ji Qiao, Xinying Wang, and Zhijian Zhang

***Abstract*—This paper proposes a new approach for online power system transient security assessment (TSA) and preventive control based on XGBoost and DC optimal power flow (DCOPF). The novelty of this proposal is that it applies the XGBoost and data selection method based on the 1-norm distance in local feature importance evaluation which can provide a certain model interpretability. The method of SMOTE+ENN is adopted for data rebalancing. The contingency-oriented XGBoost model is trained with databases generated by time domain simulations to represent the transient security constraint in the DCOPF model, which has a relatively fast speed of calculation. The transient security constrained generation rescheduling is implemented with the differential evolution algorithm, which is utilized to optimize the rescheduled generation in the preventive control. Feasibility and effectiveness of the proposed approach are demonstrated on an IEEE 39-bus test system and a 500-bus operational model for South Carolina, USA.**

***Index Terms*—DC optimal power flow (DCOPF), model interpretability, preventive control, transient security assessment (TSA), XGBoost.**

NOMENCLATURE

CPP	Central power plant.
PTS	Power transformer substation.
OP	Operating point.
DT	Decision tree.
TSA	Transient security assessment.
DCOPF	Direct-current optimal power flow.
TSC-DCOPF	Transient security constrained direct-current optimal power flow.
CS	Contingency set.
OB, PB	Overall database and prepared database.
G_CPPx	MW-output of CPP x.
PGi	MW-output of generator i.
P_x_y	MW-only power flow from bus x to bus y.
P_A_B	MW-only power flow from A to B.
S_{G}	Set of generators participating in the preventive control.

Manuscript received September 2, 2020; revised November 13, 2020; accepted December 4, 2020; date of online publication December 21, 2020; date of current version January 20, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB0905900.

S. T. Zhang (corresponding author, e-mail: richardstzhang@163.com), D. X. Zhang, J. Qiao and X. Y. Wang are with China Electric Power Research Institute, Beijing 100192, China.

Z. J. Zhang is with State Grid Beijing Electric Power Dispatching and Control Center, Beijing 100031, China.

DOI: 10.17775/CSEEJPES.2020.04780

I. INTRODUCTION

WITH the increase in scale and complication of power systems as well as the continuously growing demand for electricity, power systems are forced to operate closer to their stable operational limits. Preventive control for transient security is an important part of the three-defense lines to ensure the safe and stable operation of power systems [1], [2]. In the preventive control, the transient stability of power systems is enhanced by adopting a variety of measures to adjust the operating point (OP), e.g. adjusting the MW-output of the generators, to ensure that the power system can maintain stable operations under $N-1$ contingencies. Currently, most dispatch operations of power systems are still based on the offline operation state calculation. With the construction of phasor measurement units (PMUs) and wide-area measurement systems (WAMSs), an opportunity is provided for utilizing data-driven approaches and machine learning (ML) methods in online power system transient security assessment (TSA) and preventive control.

Preventive control for power system transient security is essentially a problem of transient security constrained optimal power flow (TSCOPF). Mathematically, it is solving a large-scale nonlinear dynamic programming problem with differential algebraic equations. Compared with the conventional optimal power flow (OPF), the difficulty of the TSCOPF lies in the additional transient stability constraints. In [4], the original semi-infinite TSCOPF problem is transformed into a conventional nonlinear optimization problem by converting the differential equations into difference equations, which can lead to dimensional disaster. As an improvement to this approach, a complicated power system with multiple generators can be reduced to a one-machine infinite-bus (OMIB) equivalent system [5]–[8]. In [9] and [10], constraints of transient energy margin based on controlling the unstable equilibrium point (CUEP) are incorporated into the TSCOPF model. In [11], the method of extended equal area criterion (EEAC) is adopted for the TSCOPF. However, the aforementioned methods have performed an equivalent process on the power system, thus the results are not always the same as that of the original system. On the other hand, the preventive control can also be implemented based on several sensitivities without an

optimization calculation [12]–[14], but these types of methods are simplistic and usually cannot achieve reasonable results in a realistic complicated power system.

With the rapid development of artificial intelligence (AI) technology in recent years [15], more and more AI algorithms have been applied in preventive control. Intelligent optimization algorithms, e.g. particle swarm optimization (PSO) [16], and the differential evolution (DE) algorithm [15], have been used in TSCOPF. In [18]–[20], artificial neural networks (ANN) are adopted to implement the transient security constraints. In [21], a systematic approach for dynamic security assessment and preventive control is proposed based on a decision tree (DT). Various methods for preventive control have been proposed based on other AI algorithms, e.g. pattern discovery (PD) [22], and support vector machine (SVM) [23]. XGBoost [24] has been widely used to achieve state-of-the-art results on many machine learning challenges. At present, XGBoost already has several applications in the field of power systems. The study in [25] applies XGBoost in the fault detection of wind turbines. In [26], XGBoost is utilized for malicious synchrophasor detection based on historical operational data. In [27] and [28], methods for TSA based on XGBoost are proposed. These applications show the potential of applying XGBoost in the prevention control for power system transient security.

Although black-box AI methods, e.g. SVM, ANN, and XGBoost, have achieved encouraging results in various areas, the lack of transparency has limited their applications under safety-critical scenarios, e.g. operation control of power system. Explainable artificial intelligence (XAI) tries to solve this problem by providing human understandable explanations [29], the research of which are generally carried out in terms of similar classification examples [30], [31], human-friendly concepts [32], or local feature importance [33], [34]. The last term is considered in this paper, which can provide a local explanation of the black-box AI model by conducting feature importance evaluation in the neighborhood of a test sample.

An approximate calculation of the MW-only power flow can be performed by the direct-current power flow (DCPF) [35]. Due to its fast calculation speed with no convergence problem, the direct-current optimal power flow (DCOPF) has been widely used in many areas of power systems [36]–[38]. The conventional DCOPF ignores the branch resistance, thus the network loss cannot be taken into account. However, for a large-scale power system, the network loss must be considered [39]. In [40], a modified DCOPF algorithm based on the network loss equivalent load model is introduced, which can achieve adequate accuracy while retaining the fast calculation speed of the conventional DCOPF method.

Based on the aforementioned researches, this paper proposes an XGBoost and DCOPF based approach for online TSA and preventive control. The OPs in the database are randomly generated based on historical generation scheduling data and the transient security results of the OPs are labeled by time domain (T-D) simulations. A data selection technique with 1-norm distance is adopted to reduce the number of training samples and help provide model interpretability, and method of SMOTE+ENN is applied for data rebalance. The DCOPF model is adopted with a fast calculation speed. The XGBoost model is introduced, which is utilized to evaluate the importance of features and represent the transient security constraint in the TSC-DCOPF model. Combined with steady-state constraints, the feasible region of the TSC-DCOPF model can be described. Then, the optimal OP is optimized by using DE and adjusted by generation rescheduling.

The remainder of this paper is organized as follows. In Section II, principles of XGBoost and formulations of the proposed TSC-DCOPF model are introduced. Section III describes the method for database preparation. Section IV presents the strategy and optimization scheme of the proposed method. Then, the proposed approach is demonstrated in Section V on an IEEE 39-bus test system and a 500-bus operational model for South Carolina, USA. The concluding remarks are provided in Section VI.

II. Methodology Description

A. Principles of XGBoost

XGBoost is an end-to-end machine learning system for tree boosting developed by Chen and Guestrin [24]. XGBoost is an ensemble learning method which conducts learning tasks by combining multiple decision trees (DTs). The DTs in XGBoost are classification and regression trees (CART) [41]. The ensemble learning methods can be roughly divided into two categories: bagging (i.e. parallel generation method of the learners) and boosting (i.e. serial generation method of the learners). A representative method of bagging is random forest (RF), while the commonly-used boosting methods are AdaBoost, GBDT, lightGBM, and XGBoost. At present, XGBoost is among the best-performing algorithms of supervised machine learning.

For a given data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i \in \mathbf{R}^m, y_i \in \mathbf{R}\}$ which has n samples and m features, the ensemble model of K DTs constructed by XGBoost is shown in Fig. 1.

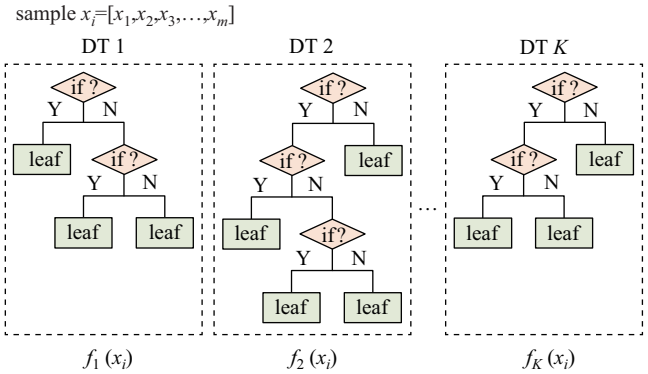

Fig. 1. The ensemble model of K DTs constructed by XGBoost. The diamonds are branch nodes and the rectangles are leaf nodes. The meaning of the branch condition "if?" is "if $\boldsymbol{x}_j < M_j$ (M_j is a real number)," e.g. "if $\boldsymbol{x}_2 < 0.35$."

According to the branch conditions, the corresponding leaf node value for a sample can be obtained in a DT. The final output value for each input sample $\boldsymbol{x}_i$ is predicted by the ensemble model of K DTs as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i),\ \ f_k \in \mathcal{F} \tag{1}$$

where $\mathcal{F}$ is the space of the DTs, $f_k(\boldsymbol{x}_i)$ is the leaf score (i.e. the prediction value) of the k-th DT for the i-th sample. The final prediction value $\widehat{y}_i$ is provided by summing up the corresponding leaf values of the K DTs. The structures and parameters of the K DTs are learned by minimizing the objective function L as follows:

$$L = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where l is the loss function for the prediction value $\widehat{y}_i$ and the target value y_i, the Ω is adopted to penalize the complexity of the DTs. The formula of Ω is as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{3}$$

where T is the number of leaves in the DT, ω_j is the score value of the leaf j, γ and λ are constants to present the degrees of regularization. The regularization term Ω can help smooth the ω to avoid over-fitting. It is worth mentioning that the shrinkage and column subsampling techniques are also adopted for the prevention of over-fitting in XGBoost.

The ensemble model of K DTs is trained through the gradient boosting process in XGBoost, as shown in Fig. 2.

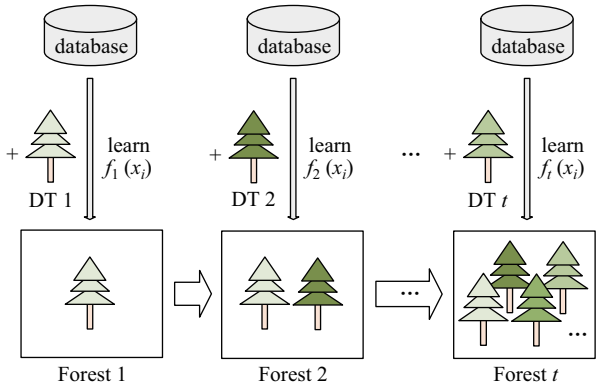

Fig. 2. Schematic diagram of the gradient boosting process in XGBoost.

In the t-th iteration of the cumulative training, the f_t is learned and added to the current forest (i.e. the ensemble model of DTs). The formula of the final prediction value in the t-th iteration is as follows:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(\boldsymbol{x}_i) = \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i) \tag{4}$$

The f_t is learned to minimize the current objective function $L^{(t)}$ as follows:

$$L^{(t)} = \sum_{i=1}^{n} l(\hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i), y_i) + \Omega(f_t) \tag{5}$$

By the second order Taylors expansion, an approximate expression of $L^{(t)}$ is as follows:

$$L^{(t)} \simeq \sum_{i=1}^{n} \left[l(\hat{y}_i^{(t-1)}, y_i) + g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \Omega(f_t) \tag{6}$$

$$g_i = \frac{\partial l(\hat{y}_i^{(t-1)}, y_i)}{\partial \hat{y}_i^{(t-1)}},\quad h_i = \frac{\partial^2 l(\hat{y}_i^{(t-1)}, y_i)}{\partial (\hat{y}_i^{(t-1)})^2} \tag{7}$$

The constant terms can be removed to obtain the simplified objective function as follows:

$$\begin{aligned} \tilde{L}^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \end{aligned} \tag{8}$$

where I_j is the index set of the samples mapping to leaf j. For a fixed DT structure, the optimal ω_j is as follows:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{9}$$

The corresponding optimal value of the current objective function is as follows:

$$\tilde{L}^{(t),*} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{10}$$

Equation (9) can be utilized to quantify the quality of a tree structure. However, it is normally impossible to check all the possible tree structures for the best one. A greedy algorithm which iteratively splits one leaf into two is used instead. For the newly generated node, all features should be tested. For each feature, a linear scan of the metric gain from the splitting is performed to obtain the best splitting position. The new node is then constructed with the best feature (i.e. feature with the largest metric gain) and the corresponding best splitting position. The formula of the metric gain is as follows:

$$\begin{aligned} G = \frac{1}{2} \Bigg[&\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \\ &\frac{\left(\sum_{i \in I} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \Bigg] - \gamma \end{aligned} \tag{11}$$

where $I = I_L \cup I_R, I_L$ and I_R are the index sets of the samples mapping to the newly generated left and right nodes respectively. In addition, besides the greedy algorithm, several approximation and variants can also be adopted to improve the effectiveness of the splitting in XGBoost.

The importance of the features can be evaluated and ranked according to the trained XGBoost model. There are three commonly-used measurements for the calculation of feature importance scores: Gain, Cover, and Frequency, among which Gain is the most relevant attribute that explains the relative importance of each feature. In this paper, the metric of total

Gain is adopted to calculate the feature importance score, which is as follows:

$$S_i = G_{\text{sum},i} \tag{12}$$

where $G_{\text{sum},i}$ is the total Gain across all the split nodes of the forest in which the i-th feature is used, and S_i is the importance score of the i-th feature. The higher value of this metric means that the corresponding feature is more important for generating prediction results than other features.

B. Proposed TSC-DCOPF Model

In this paper, a transient security constrained generation rescheduling method, based on DCOPF, is proposed to implement the online preventive control approach when the current OP of the power system is predicted to be insecure by the trained XGBoost model. The proposed TSC-DCOPF model, which includes an objective function with equality and inequality constraints, can be formulated as follows:

$$\begin{aligned} \min\ & f(\boldsymbol{u}, \boldsymbol{v}) \\ \text{s.t. } & h(\boldsymbol{u}, \boldsymbol{v}) = 0 \\ & g(\boldsymbol{u}, \boldsymbol{v}) \leqslant 0 \\ & F_{\text{TS}}(\boldsymbol{u}, \boldsymbol{v}) < 0 \end{aligned} \tag{13}$$

where $\boldsymbol{u}$ represents the controllable variables which are the MW-output of the generators, $\boldsymbol{v}$ represents the dependent variables which are constrained by power flow equations, e.g. the MW-only power flow of the branches, and $f(\boldsymbol{u}, \boldsymbol{v})$ is the objective function, which is as follows:

$$f(u) = \sum_{i \in S_{\text{G}}} \left| P_{\text{G},i} - P^0_{\text{G},i} \right| \tag{14}$$

where $P^0_{\text{G},i}$ and $P_{\text{G},i}$ are the MW-output of the i-th generator before and after generation rescheduling, and S_G is the set of generators participating in the preventive control.

The equalities $h(\boldsymbol{u}, \boldsymbol{v}) = 0$ represent the power flow equations. Practically, under the realistic scenario of online operations, the to-be-adjusted objects, which require the most attention of the operators, are the MW-only power flow of transmission lines or sections. Due to the advantages of DCOPF mentioned before, the DCPF equations are adopted as follows [39]:

$$\boldsymbol{P} = \boldsymbol{P}_{\text{gen}} - \boldsymbol{P}_{\text{load}} - \boldsymbol{P}_{\text{shunt}} - \boldsymbol{P}_{\text{loss}} - \boldsymbol{P}_{\text{hvdc}} = \boldsymbol{B}\boldsymbol{\theta} \tag{15}$$

where $\boldsymbol{P}_{\text{gen}}$ is the MW-only power vector of the generators, $\boldsymbol{P}_{\text{load}}$ is the MW-only power vector of the loads, $\boldsymbol{P}_{\text{shunt}}$ is the bus shunt loss vector, $\boldsymbol{P}_{\text{loss}}$ is the branch loss vector, $\boldsymbol{P}_{\text{hvdc}}$ is the HVDC power infeed vector of the nodes with HVDC lines, $\boldsymbol{P}$ is the node injected MW-only power vector, $\boldsymbol{\theta}$ is the node voltage phase angle vector except the slack node, and $\boldsymbol{B}$ is the node susceptance matrix which is implemented with sparse matrix technique. The $\boldsymbol{\theta}$ can be obtained by solving this linear equation. Among them, the $\boldsymbol{P}_{\text{shunt}}$ and $\boldsymbol{P}_{\text{loss}}$ need to be estimated. The approximate $P_{\text{shunt},i}$ can be calculated as follows:

$$P_{\text{shunt},i} = g_{i,0} \tag{16}$$

where $g_{i,0}$ is the total shunt conductance of bus i. The formula to estimate the $P_{\text{loss},i}$ is as follows [38]:

$$P_{\text{loss},i} = \sum_{\substack{j \in i \\ j \neq i}} \frac{P_{\text{loss},ij}}{2} \tag{17}$$

$$P_{\text{loss},ij} = |I'_{ij}|^2 r_{ij} = (\alpha_{ij} P'_{ij})^2 r_{ij} \tag{18}$$

$$P'_{ij} = \frac{\theta_i - \theta_j}{x_{ij}} \tag{19}$$

where θ_i and θ_j are the voltage phase angles of node i and j, r_{ij}, x_{ij}, I'_{ij} and P'_{ij} are the resistance, reactance, current and the MW-only power flow of the branch ij, respectively. The unit of phase angle is radians. It should be noted that the per-unit values are adopted in this paper, except the units for time or phase angle. For the unit of power, 1.0 p.u. is equivalent to 100 MW. The α_{ij} is a scale factor to estimate I'_{ij} with P'_{ij}. The $\boldsymbol{P}_{\text{loss}}$ should be calculated iteratively with (15)–(19), until the following formula is satisfied:

$$\|\boldsymbol{P}_{\text{loss},k} - \boldsymbol{P}_{\text{loss},k-1}\| \leqslant \varepsilon \tag{20}$$

where $\boldsymbol{P}_{\text{loss},k}$, $\boldsymbol{P}_{\text{loss},k-1}$ are the branch loss vector at the k-th and $(k-1)$-th iterations respectively, ε is the maximum allowable error value of the branch loss, which is taken as 10^{-4}.

The inequalities $g(\boldsymbol{u}, \boldsymbol{v}) \leqslant 0$ represent the basic operational constraints of the power system, which are as follows:

$$P_{\text{G},i,\min} \leqslant P_{\text{G},i} \leqslant P_{\text{G},i,\max} \tag{21}$$

$$-P_{\text{P},k,\max} \leqslant P_{\text{P},k} \leqslant P_{\text{P},k,\max} \tag{22}$$

where $P_{\text{G},i,\max}$ and $P_{\text{G},i,\min}$ are the maximum and minimum MW-output of the i-th generator, $P_{\text{P},k}$ and $P_{\text{P},k,\max}$ are the MW-only power flow and the maximum MW-only power flow of the k-th transmission line or section.

The inequality $F_{\text{TS}}(\boldsymbol{u}, \boldsymbol{v}) < 0$ represents the operational constraint for transient security formed by the XGBoost model, which is trained with a large number of operation samples. The system for each sample should be tested whether it can or cannot maintain transient stability under each contingency of the contingency set (CS) by applying T-D simulations. To ensure the reliability of the results, the time length of the T-D simulation is set to be 20 s. The employed evaluation criteria for transient stability are as follows:

- Transient angle stability: The system is considered to be transient angle unstable if the maximum angle separation of any two rotor angles in degree $\Delta\delta_{\max} > 360^\circ$ [22].
- Transient voltage stability: The system is considered to be transient voltage unstable if any bus voltage amplitude is unable to recover to be above 0.8 p.u. within 10 s [42].

For each sample, the controllable variables $\boldsymbol{u}$ and the dependent variables $\boldsymbol{v}$, i.e. the MW-output of the generators or CPPs (G_CPP*x*) and MW-only power flow of the transmission lines or sections (P_A_B), are combined to form the corresponding sample input data $\boldsymbol{x}$ as follows:

$$\begin{aligned} x &= [x_1, x_2, x_3, \ldots, x_M] \\ &= [\text{G\_CPP1}, \text{G\_CPP2}, \ldots, \text{P\_A\_B}, \text{P\_C\_D}, \ldots] \end{aligned} \tag{23}$$

where x_i is the i-th feature of sample input data $\boldsymbol{x}$, and M is the number of features. It can be seen that $F_{\mathrm{TS}}(\boldsymbol{u},\boldsymbol{v}) = F_{\mathrm{TS}}(\boldsymbol{x})$.

Classification of secure or insecure samples is a two-class classification problem in which the cross-entropy loss function is generally adopted as follows:

$$\mathrm{Loss} = -\left[y \cdot \log(p) + (1-y) \cdot \log(1-p)\right] \tag{24}$$

where $\log(\cdot)$ is the natural logarithm function, y is the label of a sample, $y = 0$ or 1, 0 represents secure, 1 represents insecure, and p is the predicted probability that the sample is insecure, which is calculated by the Logistic function as follows:

$$p = \mathrm{Logistic}(\hat{y}) = \frac{1}{1+e^{-\hat{y}}} \tag{25}$$

$$\hat{y} = F_{\mathrm{TS}}(\boldsymbol{x}) = \sum_{k=1}^{K} f_k(\boldsymbol{x}) \tag{26}$$

where f_k is the k-th DT in the XGBoost model. A sample is classified by comparing its predicted probability value p with the classification threshold $p_{\mathrm{th}} = 0.5$. A sample with $p < 0.5$ is a secure sample, otherwise it is insecure. This is equivalent to the criterion that a sample with $F_{\mathrm{TS}}(\boldsymbol{u},\boldsymbol{v}) = F_{\mathrm{TS}}(\boldsymbol{x}) < 0$ is a secure one, otherwise it is an insecure one.

III. Database Preparation

Training of the XGBoost model is based on an overall database (OB) prepared in advance with a large number of OPs and their T-D simulation results of transient security. These OPs include historical generation scheduling data and further randomly generated ones with steady-state constraints and feasible power flow solutions, which are as follows:

$$\mathrm{OP} = [P_{\mathrm{G},1}, P_{\mathrm{G},2}, P_{\mathrm{G},3}, \ldots, P_{\mathrm{G},N_{\mathrm{G}}}] \tag{27}$$

$$\mathrm{OP\_CPP} = [\mathrm{G\_CPP1}, \mathrm{G\_CPP2}, \ldots, \mathrm{G\_CPP}N_{\mathrm{C}}] \tag{28}$$

where N_{G} and N_{C} are the number of generators and CPPs that can be dispatched in the preventive control, respectively. For an insecure initial OP, the generation rescheduling will be carried out under its current load level. Therefore, different databases are suggested to be established with different corresponding load levels. Typical ones of $N-1$ contingencies in which transient instability occur frequently are selected to form the CS according to historical records and experience of the operators. In normal circumstances, these contingencies are the most critical 3-phase-ground faults.

A. Data Selection

In this paper, data selection is adopted to reduce the number of training samples and help provide model interpretability at the same time. The interpretability is considered to be necessary if the power system operators monitor and schedule based on the prediction of a black-box AI model. It is difficult to interpret the black-box AI model from a global perspective, and may not be of necessity. In fact, local feature importance evaluation is one of the most frequently-used methods for providing model interpretability.

The local important features obtained by evaluation represent the prediction-making basis and reasons of the black-box AI model in the neighborhood of the test sample, thereby providing a local explanation of the model. The prediction of the model is considered to be unreliable if the obtained local important features are obviously inconsistent with human experience. Take LIME [34] for example, as shown in Fig. 3, an interpretable linear model is fitted with samples in the neighborhood of the test point to perform a local feature importance evaluation.

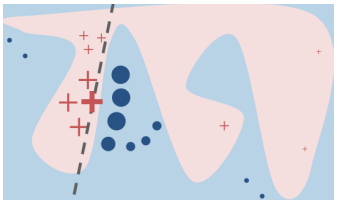

Fig. 3. Schematic diagram of LIME [32]. The dashed line approximates the decision boundary in the neighborhood of the test point (bold cross).

Practically, for a given insecure initial OP of the power system, in the process of its adjustment towards the security boundary, the nearby samples are the actually effective ones, while the far-away samples generally do not play a role. Therefore, samples close to the initial OP can be selected, as shown in Fig. 4. Then, a certain local model interpretability can be achieved by local feature importance evaluations with the XGBoost model trained with the selected samples. It should be noted that the security boundary here in Fig. 4 is only a schematic diagram with no specific expression.

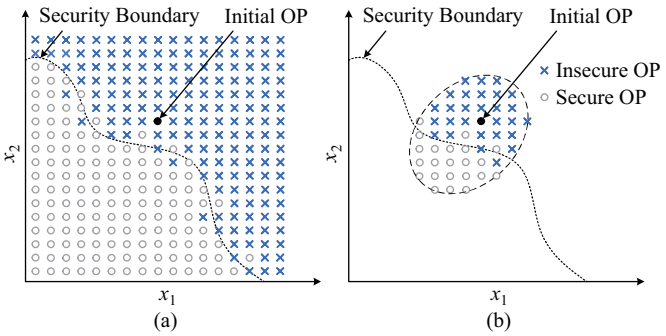

Fig. 4. Schematic diagram of data selection. (a) Samples before data selection. (b) Samples after data selection.

There are several commonly-used metrics that can describe the distance between two samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, e.g. ∞-norm distance d_∞, 2-norm distance d_2, and 1-norm distance d_1, which are defined as follows:

$$d_\infty(\boldsymbol{x}_1,\boldsymbol{x}_2) = \max\{|x_{1,i} - x_{2,i}|\} \tag{29}$$

$$d_2(\boldsymbol{x}_1,\boldsymbol{x}_2) = \left(\sum |x_{1,i} - x_{2,i}|^2\right)^{\frac{1}{2}} \tag{30}$$

$$d_1(\boldsymbol{x}_1,\boldsymbol{x}_2) = \sum |x_{1,i} - x_{2,i}| \tag{31}$$

where $x_{1,i}$ and $x_{2,i}$ are the i-th feature of sample $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. It should be noted that only features of G_CPPx are taken into account in the calculation of distance

and features of P_A_B are not considered. In most current research studies, the ∞-norm distance is utilized as follows:

$$0.9P^0_{\mathrm{G},i} \leqslant P_{\mathrm{G},i} \leqslant 1.1P^0_{\mathrm{G},i} \tag{32}$$

It can be recognized that the adjustment range reserved for each generator or CPP is quite small with this type of distance metric. In order to complete a certain amount of generation rescheduling, a large number of generators and CPPs need to be adjusted. The distance, defined by 2-norm or 1-norm, however, has a wider adjustment range for each generator with a certain total generation rescheduling amount of the system, which is similar to that in the realistic scheduling scenarios.

Typical results of data selection with ∞-norm distance, 2-norm distance, and 1-norm distance are shown in Fig. 5. Practically, the distance defined by the 2-norm or 1-norm metric have similar effects in data selection. The 1-norm distance can be understood as the total amount of generation rescheduling, while the 2-norm distance is the distance of two samples in the Euclidean space. In order to facilitate understanding, the 1-norm distance is adopted in this paper.

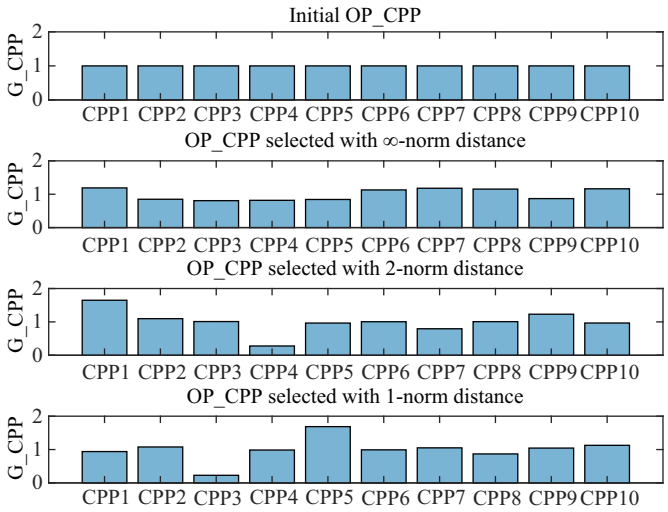

Fig. 5. Typical results of data selection with ∞-norm distance d_∞, 2-norm distance d_2, and 1-norm distance d_1. Assuming that 10 CPPs in the system could be rescheduled, each G_CPP of the initial OP_CPP is 1.0 and the total amount of generation rescheduling is 2.0.

B. Data Rebalancing

A dataset is imbalanced if instances of some classes are far fewer than those of the other classes. With an imbalanced data set, classification models of machine learning can hardly obtain satisfactory training results. However, for the TSA scenarios in this paper, an imbalanced dataset is common.

Data rebalancing aims to rebalance the proportions of different classes in an imbalanced dataset, of which random under-sampling (RUS) and random over-sampling (ROS) are the simplest methods. However, many samples are discarded in random under-sampling, which affects the training effect of the model, while random over-sampling will cause the problem of model overfitting. The synthetic minority oversampling technique (SMOTE) [43] is an improved scheme based on random oversampling. SMOTE generates new samples based on the spatial distribution of minority samples by interpolation. For each sample in the minority class, the Euclidean distance between this sample and all other minority samples is calculated to obtain its k-nearest neighbors. The difference between the original sample and a randomly chosen neighbor is multiplied by a factor with the range of 0–1, and then added to the original sample to derive a new sample.

Nonetheless, noise samples are easily generated when being inserted between marginal outliers and inliers in the previous SMOTE method. Therefore, data cleaning techniques, e.g. Tomek links, and Wilson's edited nearest neighbor rule (ENN), are suggested to be executed after oversampling, which leads to methods of SMOTE+Tomek and SMOTE+ENN [44]. In ENN, the class in which more than two neighbors of the sample belong should be marked as the predicted class. A sample should be removed if its predicted class contradicts its actual class. Generally speaking, the method of SMOTE+ENN tends to remove more noise samples than SMOTE+Tomek and is expected to provide a more in-depth data cleaning effect. Hence, the method of SMOTE+ENN is adopted for data rebalancing in this paper. Typical results of the samples with different data rebalancing techniques are shown in Fig. 6.

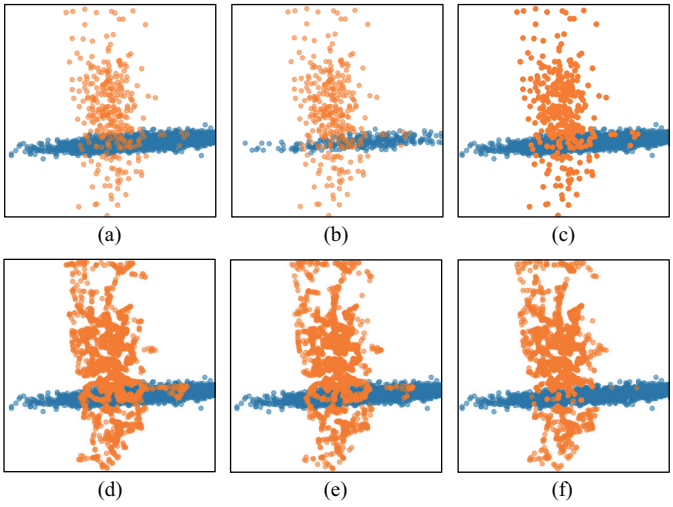

Fig. 6. Samples with different data rebalancing techniques. (a) Ordinary samples. (b) Samples after applying RUS. (c) Samples after applying ROS. (d) Samples after applying SMOTE. (e) Samples after applying SMOTE+Tomek. (f) Samples after applying SMOTE+ENN. The blue dots represent the majority samples and the orange dots represent the minority samples.

IV. Preventive Control Scheme

A. Strategy of the Preventive Control

When the power system on an initial OP is evaluated to be insecure, online preventive control actions should be conducted. In this paper, measures of generation rescheduling are considered to be adopted to restore the system from an insecure OP to a secure one in the preventive control. The flowchart of the proposed online preventive control approach is shown in Fig. 7, and the proposed methodology is implemented in the following stages.

Stage I: Preparation of the Overall Database

An overall database (OB) containing a large number of OPs, under different load levels in the system, should be prepared

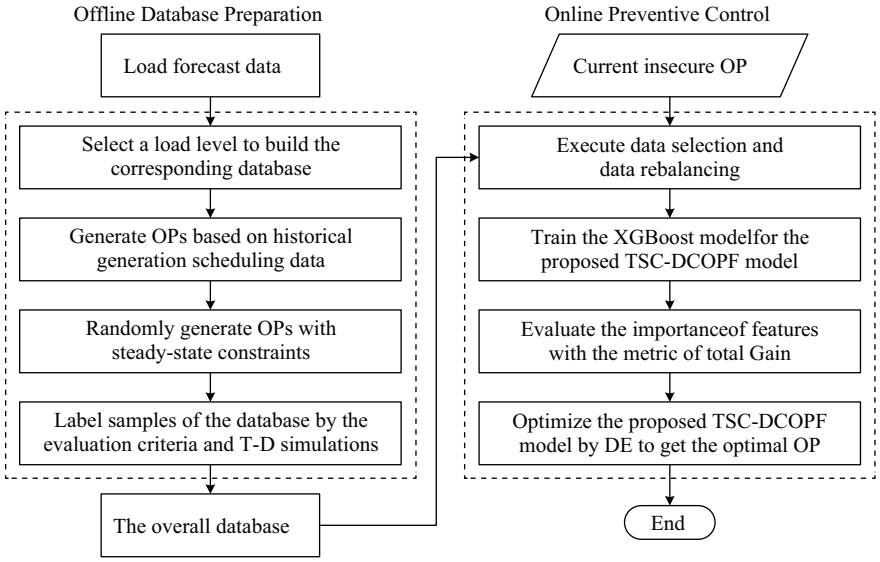

Fig. 7. Flowchart of the proposed online preventive control approach.

offline in advance. These OPs include historical generation scheduling data and further randomly generated ones with steady-state constraints and feasible power flow solutions. Each sample of the database is labeled with T-D simulations under each contingency of the CS and the evaluation criteria for transient stability. After the overall database is generated, there is no need to do T-D simulations again in the following procedures.

Stage II: Data Selection and Data Rebalancing

For a given insecure initial OP, data selection is used to reduce the scale of the database and help provide model interpretability. Since the imbalanced database is common to appear, the method of data rebalancing is adopted to rebalance the proportions of different classes in the database. Consequently, the prepared database (PB) for a given initial OP is formed.

Stage III: Training of the XGBoost Model

The XGBoost model is trained based on the generated PB to conduct the TSA and form the transient security constraint in the proposed TSC-DCOPF model. Meanwhile, the importance of features can be evaluated, ranked and selected by the XGBoost model with the metric of total Gain. The XGBoost-based TSA and preventive control scheme is shown in Fig. 8.

Stage IV: Implementation of the Preventive Control

On the basis of the proposed TSC-DCOPF model, the optimal OP can be obtained by using DE and providing it to the operators to execute the corresponding online preventive control measures when the current OP is predicted to be insecure.

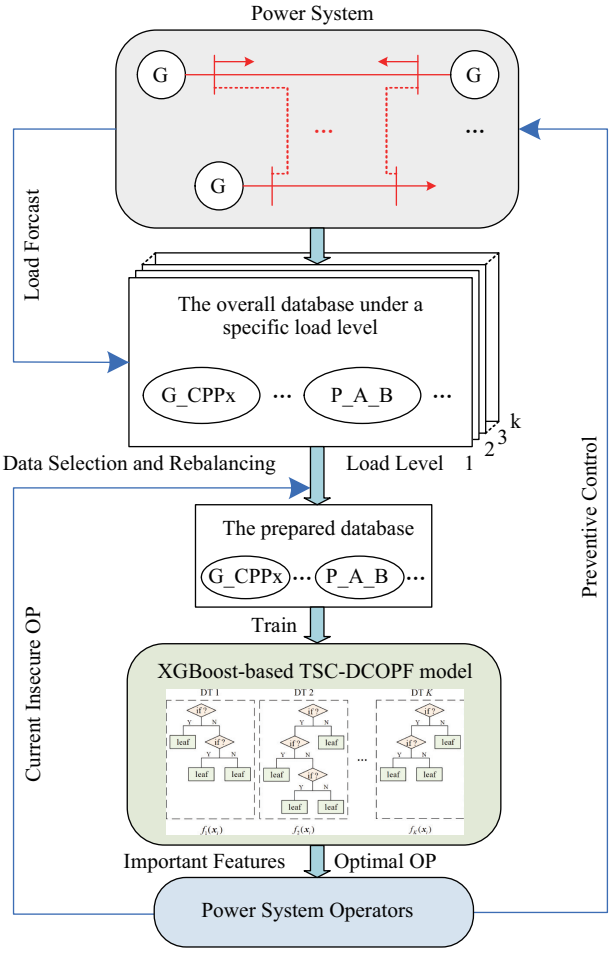

Fig. 8. XGBoost-based TSA and preventive control scheme.

B. Optimization of the Preventive Control

The differential evolution algorithm (DE) is an efficient heuristic global optimization algorithm developed by Storn and Price [45]. It is a type of swarm intelligent evolution algorithm with a simple structure to implement. In each generation of DE, new individuals are generated through differential operations and crossover operations. The older individual will be replaced if the fitness of a new individual is better. Through continuous evolution, the individuals will move toward the optimal solution. Compared with other algorithms, better op-

timization results can usually be obtained by DE when solving complex optimization problems with several constraints.

The TSC-feasible operating region is provided by the trained XGBoost model. Then, utilizing DE, the optimal OP can be searched out by minimizing the total amount of generation rescheduling, as defined in (14). Each individual of the population generated in each generation should be inside the feasible operating region with steady-state constraints, or this individual should be re-generated. The transient security constraint in the proposed TSC-DCOPF model is implemented by using penalty functions.

In fact, the main time consumption of the TSCOPF solved by DE lies in the power flow calculation. The calculation speed of the DC power flow with branch loss is generally 4 times more than that of the conventional DC power flow due to the requirement of an iterative solution for branch loss. Under the scenario of preventive control, the operational state of the system is adjusted in the neighboring area of the initial OP, thus the change of branch network loss is small and can be considered to be unchanging in the early and middle stages of generations in DE.

V. Case Studies

Experimental results of the proposed approach are presented and discussed in this section. Two test systems are selected to demonstrate the proposed TSC-DCOPF model, namely, the IEEE 39-bus test system and the South Carolina test system. The latter shows that the presented approach can be applied to realistic power systems. For both test systems, all of the generators are assumed to be available for generation rescheduling as part of the preventive control. In fact, for an insecure initial OP, the database with the corresponding load level should be selected first. To simplify the results and without loss of generality, the overall database is obtained under a typical load level with a small range of 95%-105% in both test systems. The contingencies of the CS are 3-phase-ground faults on the lines which are cleared after 0.1 s by a single line tripping. It should be noted that in each line fault contingency, faults on both sides of the line are adopted respectively.

The programs for the test cases are developed based on Python and run on a computer with an Intel Core i5-8300H 2.30 GHz CPU and 16 G RAM.

A. IEEE 39-Bus Test System

The IEEE 39-bus test system represents the 345 kV power network of New England, USA [46]. It has 39 buses, 10 generators, 12 transformers, and 34 lines, in which the No. 39 generator is an equivalent machine. The generators and loads are modeled with detailed sub-transient models and constant impedance load models, respectively. The system is manually divided into three regions according to the connection structure of the system, as shown in Fig. 11. The contingency set of the test system in this case is shown in Table I.

In fact, randomly generating new OPs with constraints are not as easy as it seems. The following simple method is proposed to implement the constrained random generation of samples.

TABLE I
Contingency Set of IEEE 39-Bus Test System

Contingency No.	Faulted Line	Faulted Bus
1	2–3	2 or 3
2	3–4	3 or 4
3	3–18	3 or 18
4	4–5	4 or 5
5	4–14	4 or 14
6	5–6	5 or 6
7	16–17	16 or 17
8	16–21	16 or 21
9	16–24	16 or 24
10	17–27	17 or 27

First, an OP should be randomly generated within the MW-output range of each generator, which is as follows:

$$\mathrm{OP} = [P_{\mathrm{G},30}, P_{\mathrm{G},31}, P_{\mathrm{G},32}, \ldots, P_{\mathrm{G},39}] \tag{33}$$

In order to maintain the balance of active power generation and consumption in the system, the following equation should be respected:

$$P_{\mathrm{G},30} + P_{\mathrm{G},31} + P_{\mathrm{G},32} + \ldots + P_{\mathrm{G},39} \approx Const. \tag{34}$$

Specifically, this equation is utilized as follows:

$$P_{\mathrm{G,sum,min}} \leqslant P_{\mathrm{G},30} + P_{\mathrm{G},31} + P_{\mathrm{G},32} + \ldots + \mathrm{P_{G,39}} \leqslant P_{\mathrm{G,sum,max}} \tag{35}$$

where $P_{\mathrm{G,sum,max}}$ and $P_{\mathrm{G,sum,min}}$ are the maximum and minimum set value of the total MW-output of the generators that can be dispatched in the preventive control. $P_{\mathrm{G\ sum,max}}$ and $P_{\mathrm{G,sum,min}}$ are two near values and the difference between them is the MW-output tolerance, which can be compensated by the MW-output change of the slack generator. Then, the previously generated OP should be adjusted in the following way if it does not satisfy (35): If the total power generation exceeds the upper limit, a generator that does not reach its lower limit of MW-output is randomly selected, and its MW-output is reduced with a small random step, e.g. 0–0.1 p.u. Similarly, if the total MW-output is below the lower limit, a generator that does not reach its upper limit of MW-output is randomly selected, and its MW-output is increased with a small random step. The above two steps are iterated repeatedly until the current OP satisfies (35). Furthermore, the generated OP should be verified whether it satisfies other steady-state constraints and has a feasible power flow solution, if not, the generated OP should be abandoned and re-generated.

For the IEEE 39-bus test system, each generator can be simply regarded as a CPP. Assuming that the MW-output range of each generator is 10%–150% of its original value, 3,000 randomly generated samples are produced to form the OB at the current load level. Then, each sample of the database is labeled by the evaluation criteria for transient stability with T-D simulations under each contingency of the CS. To facilitate elaboration, the OP in the original calculation data of the IEEE 39-bus test system is selected as the initial OP, which is an insecure OP. Then, the PB can be obtained following the procedure of data selection and data rebalancing as mentioned before. Specifically, in the data selection procedure, samples within the 1-norm distance of 12 p.u. from the initial OP are

selected from the OB. The sizes of OB and the example PB in this test system are shown in Table II.

TABLE II
SIZE OF OB AND THE EXAMPLE PB IN THE IEEE 39-BUS TEST SYSTEM

Size of Database	Secure	Insecure	Total
OB	427	2573	3000
PB	478	362	840

Afterwards, the XGBoost model is trained based on the generated PB to conduct the TSA, which achieves an average test accuracy ratio of 97.88% with a 10-fold cross validation. To facilitate visualization, the max-depth of each DT is limited to 2. The number of training rounds is set to 100. After the training procedure, 100 DTs are generated in the XGBoost model, where each sample can obtain a corresponding leaf value from each DT, as shown in Fig. 9. With the Logistic function, a sample is predicted to be secure if the sum of the corresponding leaf values in DTs is a positive value.

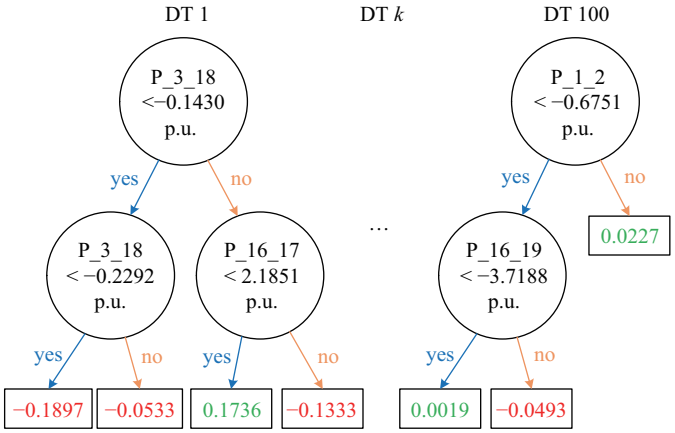

Fig. 9. DTs generated by the XGBoost model in the IEEE 39-bus test system.

The feature importance scores are calculated and ranked based on the metric of total Gain, and the top 10 are shown in Fig. 10 and Fig. 11. It can be recognized that the most important features are the MW-only power flow of the lines rather than the MW-output of the generators. In general, the MW-only power flow transmission from Region-1 to Region-2 and Region-3 through lines 16–17 has an important impact on the transient security of the IEEE 39-bus test system on the

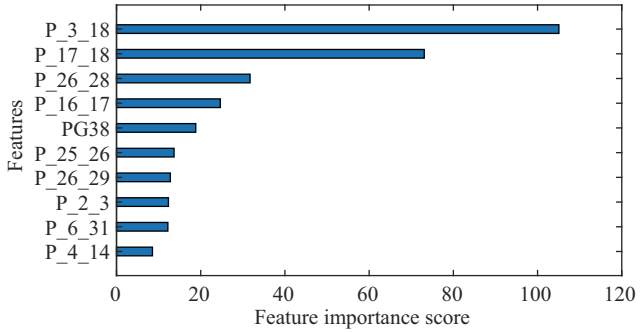

Fig. 10. 599531495The top 10 feature importance scores based on the metric of total Gain in the IEEE 39-bus test system on the initial OP.

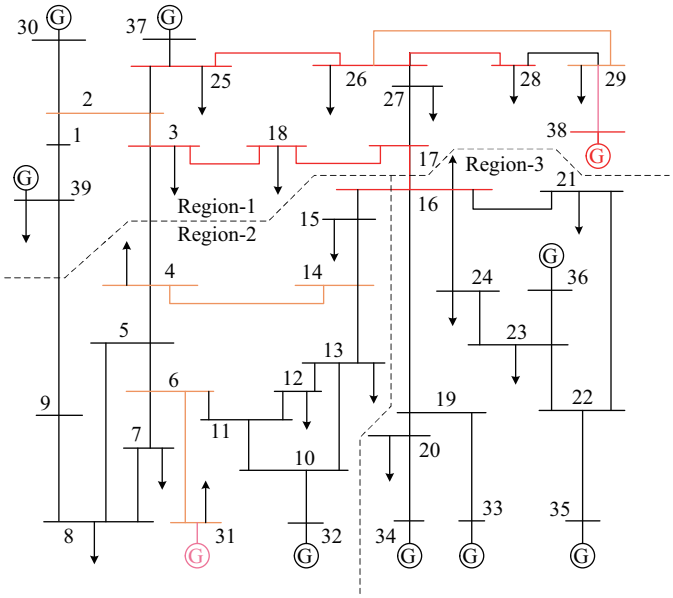

Fig. 11. The top 10 important features based on the metric of total Gain in the IEEE 39-bus test system. The top 6 features are drawn in red, the other 4 are drawn in orange, and their directly related features are drawn in pink.

example initial OP.

Next, the ranked features are selected one by one from top to bottom and added to the selected features to train the XGBoost models. The accuracy curve with the number of selected features is shown in Fig. 12, where the accuracy tends to be stable when the number of selected features reaches 6. Hence, the top 6 features can be selected to train the XGBoost model and form the transient security constraint in the proposed TSC-DCOPF model.

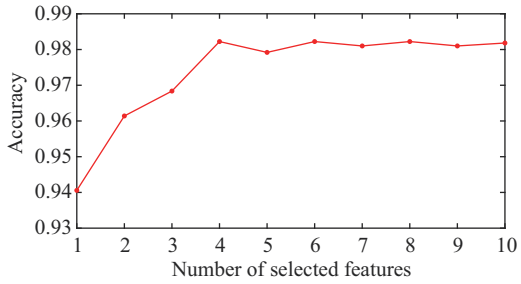

Fig. 12. Accuracy with the number of selected features based on the metric of total Gain in the IEEE 39-bus test system on the initial OP.

In order to verify the proposed model interpretation method in terms of local feature importance, LIME is utilized to explain the XGBoost model trained with the OB, where the initial insecure OP is the test sample. The top 10 important features provided by LIME are shown in Fig. 13. It can be seen that the most important features provided by LIME, e.g. P_16_17, P_3_18, PG38, are consistent with those provided by the proposed model interpretation method in this paper.

Then, the proposed TSC-DCOPF model is optimized with DE, as shown in Fig. 14. The optimal OP is obtained after 150 generations of optimization in 5.51 s, where the population size of each generation is 10. The MW-output of generators

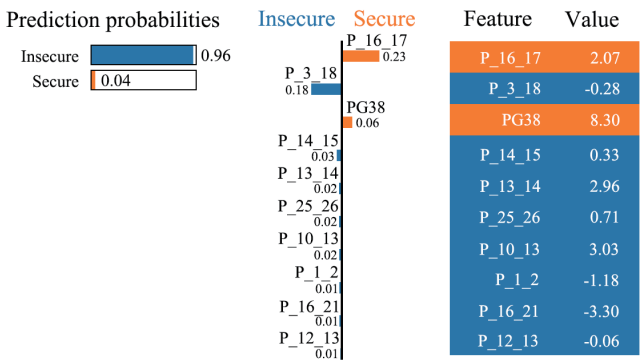

Fig. 13. The top 10 important features provided by LIME in IEEE 39-bus test system on the initial insecure OP.

before and after optimization is shown in Fig. 15. The MW-outputs with relatively large variations are PG34, PG36, and PG37, as shown in Table III. In fact, if the other generators remain unchanged, the system can still restore security as long as these three generators are adjusted with these optimized values. On the selected insecure initial OP in this case, the system is transient unstable when a 3-phase-ground fault occurs on the side of bus 16 in line 16–17. Under this contingency, the rotor angle curves of the generators with respect to the center of inertia (COI) and voltage amplitudes of the buses before and after generation rescheduling are shown in Fig. 16 and Fig. 17. For the convenience of observation, only a portion of the 0-20 s T-D simulation results are presented in the figures. It can be recognized that the IEEE 39-bus test system for the example insecure initial OP has restored transient security with the generation rescheduling for preventive control.

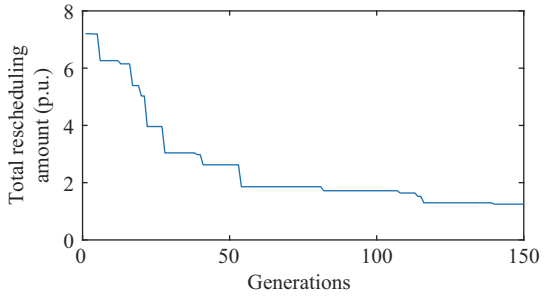

Fig. 14. Convergence curve of DE for the proposed TSC-DCOPF model in IEEE 39-bus test system.

TABLE III
GENERATION RESCHEDULING RESULT IN THE IEEE 39-BUS TEST SYSTEM

MW-output (p.u.)	PG34	PG36	PG37
Before rescheduling	5.0800	5.6000	5.4000
After rescheduling	4.7260	5.3496	5.9416
Changed	−0.3540	−0.2504	0.5416

B. South Carolina Test System

The South Carolina test system is a 500-bus operational model jn South Carolina, USA, which is built using the statistical analysis of a real power system. It has 500 buses, 90 generators, 131 transformers, and 295 lines. It has 208 stations in total, including 31 CPPs and 177 PTSs, as shown in Fig. 22. Among them, 49 generators with a capacity larger than

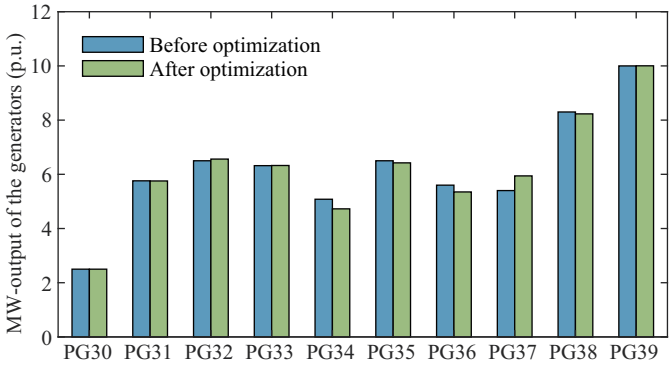

Fig. 15. MW-output of the generators before and after optimization in IEEE 39-bus test system on the initial OP.

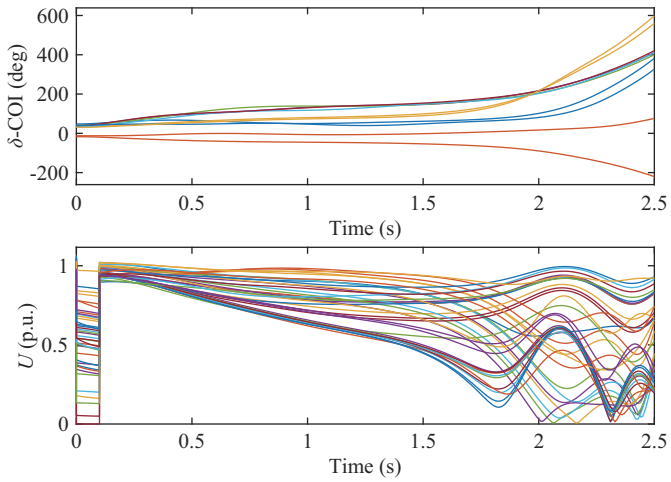

Fig. 16. Rotor angle curves of the generators and voltage amplitudes of the buses before generation rescheduling with a 3-phase-ground fault occurred on the side of bus 16 of line 16-17 in the IEEE 39-bus test system on the initial OP.

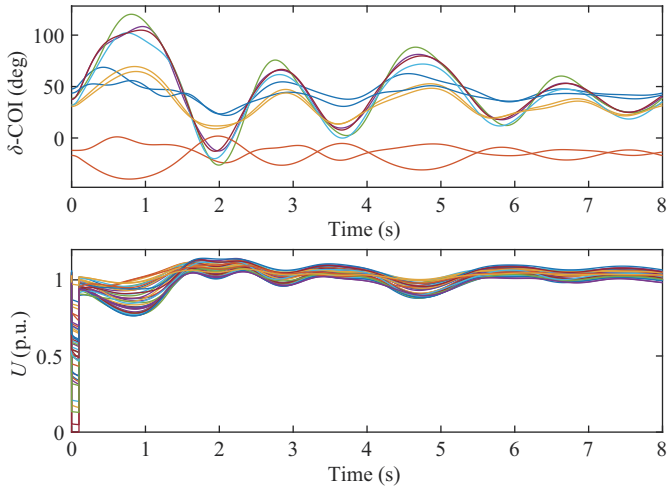

Fig. 17. Rotor angle curves of the generators and voltage amplitudes of the buses after generation rescheduling with a 3-phase-ground fault occurred on the side of bus 16 of line 16-17 in the IEEE 39-bus test system.

50 MW in 19 CPPs are selected to be adjusted in this case. The system is manually divided into three regions to facilitate the visualization of the OPs in the databases. The generators in the system are modeled with detailed sub-transient models and the loads are modeled with comprehensive load models, including a 40% constant impedance and 60% induction motor. The parameters of the induction motor model are typical parameters, among which the stator leakage reactance is 0.295 p.u. and the rotor leakage is 0.12 p.u. The contingency set of the South Carolina test system in this case is shown in Table IV.

TABLE IV
CONTINGENCY SET OF SOUTH CAROLINA TEST SYSTEM

Contingency No.	Transmission Channel	Faulted Line	Faulted Bus
1	CPP1-PTS44	3–62	3 or 62
2	CPP24-PTS26	232–262	232 or 262
3	PTS35-PTS37	163–200	163 or 200
4	PTS35-CPP50	162–220	162 or 220
5	PTS55-PTS56	87–141	87 or 141
6	PTS58-CPP64	143–452	143 or 452
7	CPP64-PTS65	143–401	143 or 401
8	PTS68-CPP69	110–274	110 or 274
9	CPP133-CPP135	14–386	14 or 386
10	CPP190-CPP197	80–407	80 or 407

In this test case, the OPs of the system are generated with a combination of switching on and switching off generators as well as adjusting the MW-output of the generators. For the generation of the OB, the OP_CPPs are first randomly generated within the MW-output range of each CPPs while considering the constraint of maintaining the balance of active power generation and consumption in the system, which are defined as follows:

$$\text{OP\_CPP} = [\text{G\_CPP1}, \text{G\_CPP3}, \ldots, \text{G\_CPP197}] \quad (36)$$

$$\text{G\_CPP1} + \text{G\_CPP3} + \ldots + \text{G\_CPP197} \approx \text{Const.} \quad (37)$$

Then, the previously generated OP_CPP should be adjusted following the similar steps described in Section V.A until it satisfies (37). However, the OP_CPPs need to be further dispatched to be OPs to perform power flow analysis and T-D simulations. Specifically, the OP_CPPs are allocated to the OPs in the following way: Priority should be given to switching on the generators with large rated MW-output values and the number of switched on generators should be as few as possible. Next, the MW-output of the switched on generators need to be adjusted to fit the corresponding value in the generated OP_CPP, where the generators with smaller rated MW-output values should be considered first. In addition, at least one generator should be kept on in each CPP. By this means, 6,000 randomly generated OPs are produced to form the OB at the current load level, which are shown in Fig. 18.

To facilitate elaboration, an insecure initial OP is randomly generated and then the corresponding PB can be obtained following the aforementioned procedure of data selection and data rebalancing. Specifically, in the data selection, samples within the 1-norm distance of 27 p.u. from the initial OP are selected from the OB. Visualization of the OPs in the PB are shown in Fig. 19. Sizes of OB and the example PB in this test system are shown in Table V.

Afterwards, the XGBoost model is trained based on the generated PB, which achieves an average test accuracy ratio

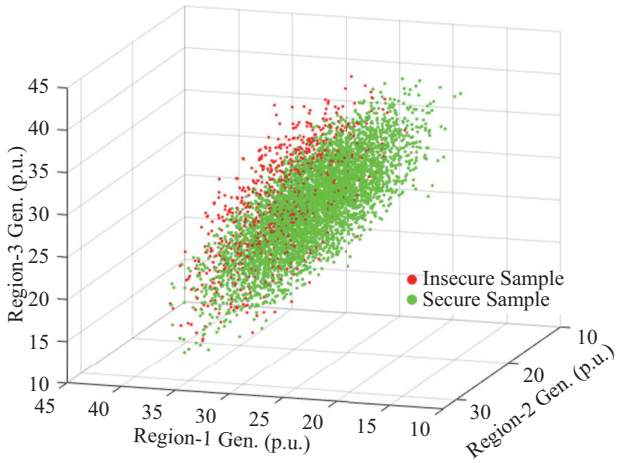

Fig. 18. Sampled OPs of the OB in South Carolina test system.

TABLE V
SIZE OF OB AND THE EXAMPLE PB IN SOUTH CAROLINA TEST SYSTEM

Size of Database	Secure	Insecure	Total
OB	5294	706	6000
PB	483	405	888

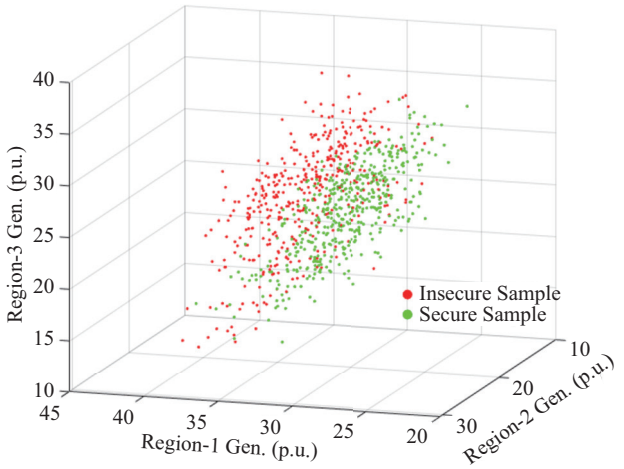

Fig. 19. Sampled OPs of the example PB in South Carolina test system.

of 98.62% with the 10-fold cross validation. The max-depth of each DT is limited to 4 and the number of training rounds is set to 100. The 100 DTs generated within 100 training rounds in the XGBoost model are shown in Fig. 20.

Then, the feature importance scores are calculated and ranked based on the metric of total Gain, and the top 10 are shown in Fig. 21 and Fig. 22. It can be recognized that the MW-only power flow transmission between Region-1 and Region-2 has an important impact on the transient security of the South Carolina test system.

The top 10 important features provided by LIME to explain the XGBoost model trained with the OB are shown in Fig. 23, where the initial insecure OP is taken as the test sample. It can be seen that the most important features obtained by LIME are consistent with those evaluated and ranked before, e.g. P_PTS36_PTS34 or P_PTS36_PTS38 (which are directly related features), P_PTS51_PTS46, P_PTS35_CPP50,

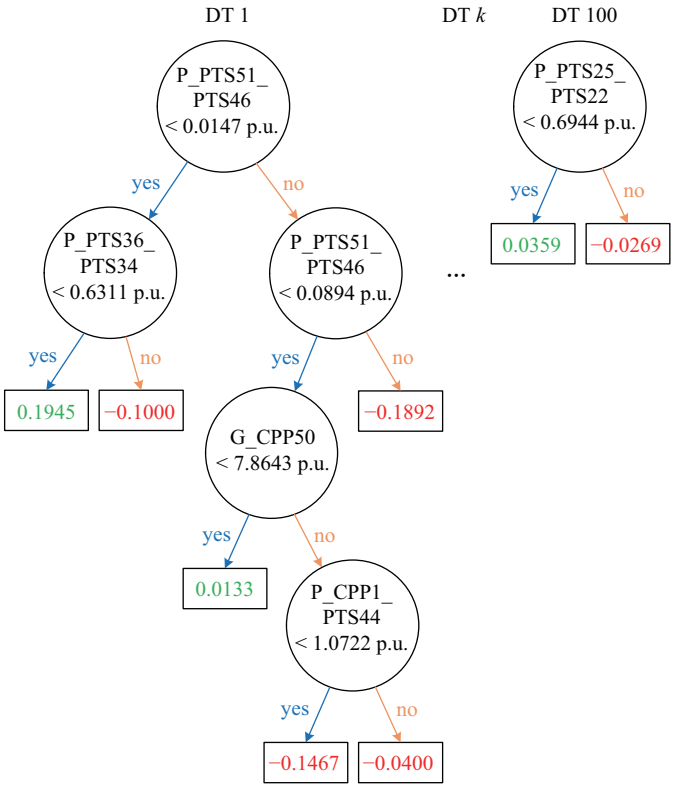

Fig. 20. DTs generated by the XGBoost model in South Carolina test system.

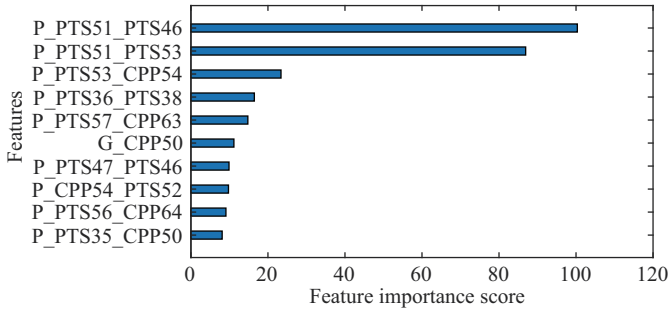

Fig. 21. The top 10 feature importance scores based on the metric of total Gain in South Carolina test system on the initial OP.

and P_PTS56_CPP64. Meanwhile, in both the IEEE 39-bus test case and South Carolina test case, most transmission lines with high feature importance scores are in or near the fault lines of CS. In fact, the fault lines of CS are the lines that often lead to transient instability of the system with the corresponding contingencies, selected by the N-1 screening or the experience of operators. As can be seen, the MW-only power flow of transmission lines in or near the fault lines of CS have a superior impact on the transient security of the test systems, which is also consistent with the experience of the operators.

Then, the XGBoost model is trained with the top 10 features after the procedure of feature selection mentioned in Section V.A. The CPPs near the important transmission lines are selected to be rescheduled, for they are generally the most effective CPPs to adjust the power flow of these lines, which are CPP3, CPP4, CPP13, CPP33, CPP43, CPP50, CPP63, CPP64, CPP69, and CPP71. Then, the proposed TSC-DCOPF model is optimized with DE. The optimal OP is obtained after 150 generations of optimization in 24.12 s, where the population size is 10. The MW-output of the selected CPPs before and after optimization is shown in Fig. 24. The MW-output with relatively large variations are G_CPP3, G_CPP13, G_CPP43, and G_CPP64, as shown in Table VI.

TABLE VI
GENERATION RESCHEDULING RESULT IN SOUTH CAROLINA TEST SYSTEM

MW-output (p.u.)	G_CPP3	G_CPP13	G_CPP43	G_CPP64
Before rescheduling	0.7151	1.2835	0.4232	11.6241
After rescheduling	0.9321	1.7458	0.5442	10.7221
Changed	0.2170	0.4623	0.1209	−0.9020

On the example of the initial OP in this case, the South Carolina test system is transient unstable when a 3-phase-ground fault occurs on the side of bus 87 in line 87–141. Under this contingency, the rotor angle curves of the generators with respect to the center of inertia (COI) and voltage amplitudes of the buses before and after generation rescheduling are shown in Fig. 25 and Fig. 26. It can be seen that the South Carolina test system on the example of the initial OP has restored security with the generation rescheduling for preventive control.

To evaluate the overall reliability of the proposed approach, 5 OBs of different load levels from 90% to 110% with intervals of 5% are created for more experiments. On each load level, 20 insecure initial OPs are randomly generated, 99% of which are successfully adjusted back to the realistic secure region (under T-D simulations) by generation rescheduling with the proposed TSC-DCOPF model. In fact, a 100% success rate cannot be guaranteed without verification by T-D simulations which cost too much time. Nevertheless, the proposed online preventive control method can still be provided to the operators as an auxiliary tool besides offline methods.

Comparisons are made between XGBoost and other machine learning methods for TSA in the proposed approach, e.g. decision tree (DT), random forest (RF), support vector machines (SVM), multi-layer perceptron (MLP), and 1-dimensional convolutional neural network (CNN-1D). Among them, both of the XGBoost and RF are trained with 100 DTs which are CART. The MLP models are set up with one middle layer which has 500 neurons and the CNN-1D model with one 1-d convolution layer which has 50 channels and the size of the convolution kernel is 3. The activation functions are ReLU. Along with the Accuracy, the metric of the F1-Score is considered to fairly evaluate the performances of the models, which is a comprehensive measurement of the Precision and the Recall. Considering the reliability of the results, 5 initial OPs are randomly selected and the corresponding PBs are generated for 5 independent test cases. Performances of different machine learning models on the validation sets with 10-fold cross validation are shown in Table VII.

It can be recognized that the XGBoost model performs better than other models by the indexes of the Accuracy and F1-Score, but the variation is not significant. In fact, each of these machine learning methods is considered to have enough high precision to carry out TSA in the preventive control,

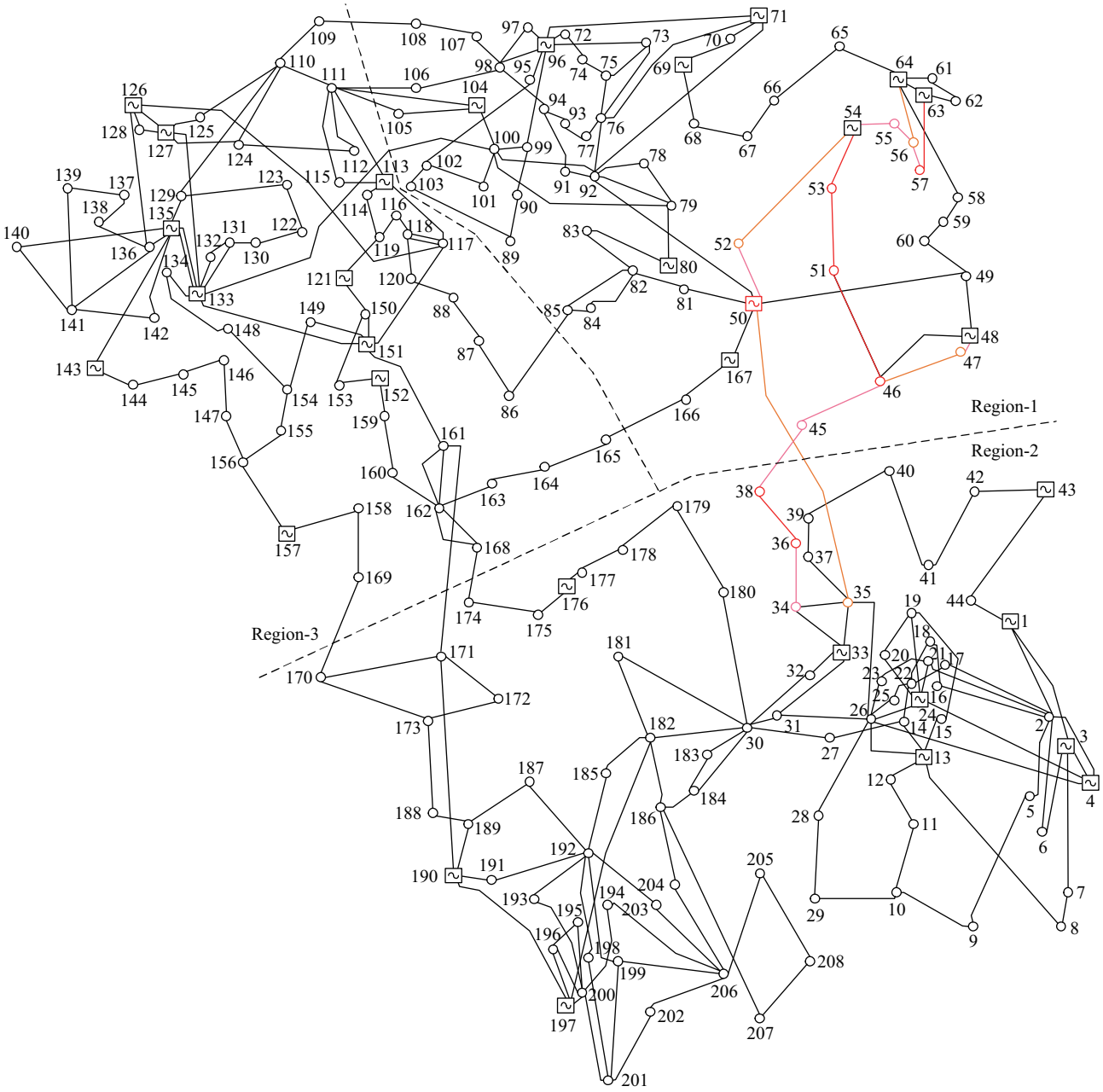

Fig. 22. The top 10 important features based on the metric of total Gain in the South Carolina test system. The top 6 features are drawn in red, the other 4 are drawn in orange, and their directly related features are drawn in pink. The number drawn in the figure is the number of stations, not the number of buses.

TABLE VII
PERFORMANCES OF DIFFERENT ML MODELS ON THE VALIDATION SETS

No.	Meas.	SVM	MLP	CNN-1D	DT	RF	XGB.
1	Acc. (%)	97.75	97.17	97.14	97.74	98.31	98.88
	F1%	98.11	97.15	97.22	97.72	98.28	99.01
2	Acc. (%)	98.32	98.09	98.18	98.03	98.03	98.32
	F1%	98.52	98.13	98.16	98.04	98.14	98.28
3	Acc. (%)	98.50	98.01	97.91	97.97	97.93	98.87
	F1%	98.70	97.97	97.86	97.99	98.12	98.85
4	Acc. (%)	98.03	98.06	98.04	97.62	98.17	98.65
	F1%	98.36	98.10	98.11	97.55	98.31	98.67
5	Acc. (%)	97.97	98.12	98.08	97.18	98.31	98.69
	F1%	98.27	98.09	98.17	97.11	98.44	98.73

as have already been implemented in many research studies. Meanwhile, compared with TSCOPF based on AC power flow, the proposed TSC-DCOPF model has a faster calculation speed, which is important when it is utilized in a realistic large-scale power system, for example, a 2507-bus test system of Northeast China, as shown in Table VIII. In the procedure of DE on each test system, 150 generations of optimization is carried out and the population size is 10.

TABLE VIII
SOLUTION TIME OF TSCOPF WITH AC POWER FLOW AND TSC-DCOPF

Time in Different Test System	39-Bus	500-Bus	2507-Bus
TSCOPF with AC Power Flow (s)	65.95	108.57	377.21
Proposed TSC-DCOPF (s)	5.51	24.12	112.41
Time Reduced (%)	91.65	77.78	70.20

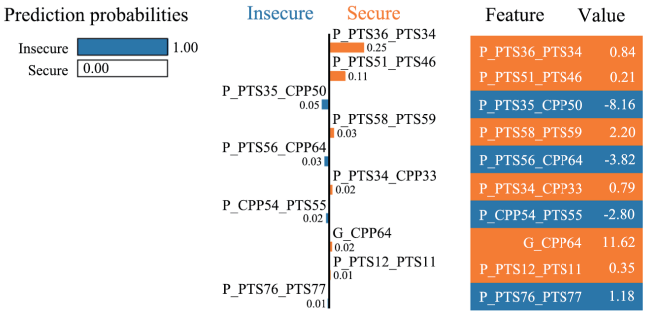

Fig. 23. The top 10 important features provided by LIME in the South Carolina test system on the initial insecure OP.

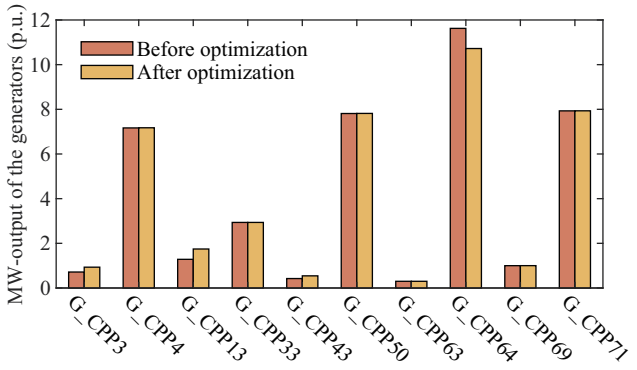

Fig. 24. MW-output of the selected CPPs before and after optimization in South Carolina test system on the initial OP.

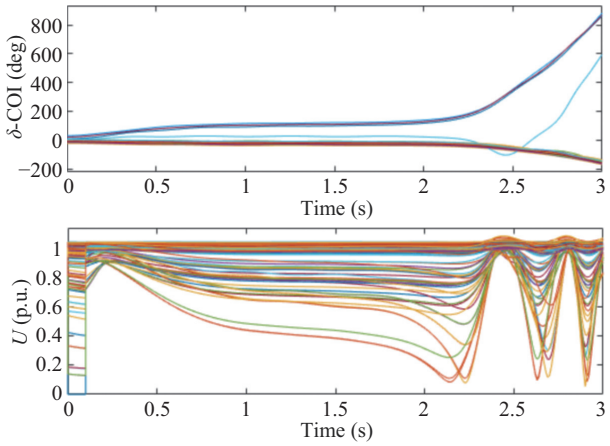

Fig. 25. Rotor angle curves of the generators and voltage amplitudes of the buses before generation rescheduling with a 3-phase-ground fault occurring on the side of bus 87 of line 87-141 in South Carolina test system on the initial OP.

Since the calculation time is sufficiently reduced with the proposed TSCOPF model compared with the conventional TSCOPF with AC power flow, the optimal OP can be obtained within tens of seconds, which could satisfy the requirement of online preventive control. For a large-scale power system, the method of network equivalent is suggested to be adopted to further increase the calculation speed.

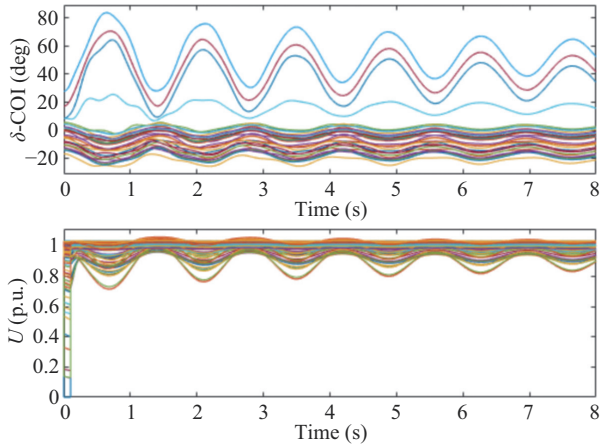

Fig. 26. Rotor angle curves of the generators and voltage amplitudes of the buses after generation rescheduling with a 3-phase-ground fault occurring on the side of bus 87 of line 87-141 in South Carolina test system.

VI. Conclusion

This paper presents a new approach for online TSA and preventive control based on XGBoost and DCOPF. The XGBoost model is utilized for local feature importance evaluation and construction of the transient security constraint in the proposed TSC-DCOPF model which is optimized by using DE. The methods of 1-norm distance and SMOTE+ENN are adopted for data selection and data rebalancing, respectively. The proposed systematic approach is demonstrated on an IEEE 39-bus test system and a 500-bus operational model in South Carolina, USA. Verified results have shown that a system on an insecure OP can be reliably adjusted to the secure region by generation rescheduling with the assistance of the proposed approach. Comparisons with other commonly-used methods indicate that the proposed approach is relatively fast and reliable with a certain model interpretability, which could meet the requirements of the engineering application for online preventive control.

In this paper, an online preventive control method of generation rescheduling has been proposed based on XGBoost and DCOPF. The major contribution of the proposed method lies in the data selection and the local feature importance evaluation which can provide the operators with the most important transmission lines or sections and CPPs for the transient security of the system on the current OP, as well as the assistance for generation rescheduling in the preventive control. A future study will focus on the method of network equivalent, selecting the generators or CPPs to be adjusted, and the further exploration of model interpretability, to further improve the practicability of the proposed method.

References

[1] Y. S. Xue, "Coordinations of preventive control and emergency control for transient stability," *Automation of Electric Power Systems*, vol. 26, no. 4, pp. 1–4, 9, Feb. 2002.

[2] Y. Tang, "Framework of comprehensive defense architecture for power system security and stability," *Power System Technology*, vol. 36, no. 8, pp. 1–5, Aug. 2012.

[3] Y. Tang, Y. H. Huang, H. Z. Wang, C. Wang, Q. Guo, and W. Yao, "Framework for artificial intelligence analysis in large-scale power grids based on digital simulation," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 4, pp. 459–468, Dec. 2018.

[4] D. Gan, R. J. Thomas, and R. D. Zimmerman, "Stability-constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 15, no. 2, pp. 535–540, May 2000.

[5] R. Zarate-Minano, T. Van Cutsem, F. Milano, and A. J. Conejo, "Securing transient stability using time-domain simulations within an optimal power flow," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 243–253, Feb. 2010.

[6] A. Pizano-Martianez, C. R. Fuerte-Esquivel, and D. Ruiz-Vega, "Global transient stability-constrained optimal power flow using an OMIB reference trajectory," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 392–403, Feb. 2010.

[7] A. Pizano-Martinez, C. R. Fuerte-Esquivel, and D. Ruiz-Vega, "A new practical approach to transient stability-constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 1686–1696, Aug. 2011.

[8] D. Ruiz-Vega and M. Pavella, "A comprehensive approach to transient stability control. I. near optimal preventive control," *IEEE Transactions on Power Systems*, vol. 18, no. 4, pp. 1446–1453, Nov. 2003.

[9] D. M. Xia, S. W. Mei, C. Shen, and A. C. Xue, "Computation of optimal power flow with transient stability margin index," *Automation of Electric Power Systems*, vol. 30, no. 24, pp. 5–10, Dec. 2006.

[10] M. B. Liu and Z. Yang, "Optimal power flow calculation with transient energy margin constraints under multi-contingency conditions," *Proceedings of the CSEE*, vol. 27, no. 34, pp. 12–18, Dec. 2007.

[11] Q. Lan, Y. J. Fang, Y. H. Bao, W. Li, T. S. Xu, and Y. S. Xue, "Transient security constrained optimal power flow based on EEAC method," *Automation of Electric Power Systems*, vol. 34, no. 8, pp. 34–38, 115, Apr. 2010.

[12] T. B. Nguyen and M. A. Pai, "Dynamic security-constrained rescheduling of power systems using trajectory sensitivities," *IEEE Transactions on Power Systems*, vol. 18, no. 2, pp. 848–854, May 2003.

[13] J, Q. Sun, D. Z. Fang, and B. R. Zhou, "Study on preventive control algorithm for dynamic security of power systems based on trajectory sensitivity method," *Power System Technology*, vol. 28, no. 21, pp. 26–30, Nov. 2004.

[14] K. N. Shubhanga and A. M. Kulkarni, "Stability-constrained generation rescheduling using energy margin sensitivities," *IEEE Transactions on Power Systems*, vol. 19, no. 3, pp. 1402–1413, Aug. 2004.

[15] D. X. Zhang, X. Q. Han, and C. Y. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[16] N. Mo, Z. Y. Zou, K. W. Chan, and T. Y. G. Pong, "Transient stability constrained optimal power flow using particle swarm optimisation," *IET Generation, Transmission & Distribution*, vol. 1, no. 3, pp. 476–483, May 2007.

[17] H. R. Cai, C. Y. Chung, and K. P. Wong, "Application of differential evolution algorithm for transient stability constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 23, no. 2, pp. 719–728, May 2008.

[18] V. J. Gutierrez-Martinez, C. A. Cañizares, C. R. Fuerte-Esquivel, A. Pizano-Martinez, and X. P. Gu, "Neural-network security-boundary constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 63–72, Feb. 2011.

[19] C. F. Kucuktezcan and V. M. I. Genc, "A new dynamic security enhancement method via genetic algorithms integrated with neural network based tools," *Electric Power Systems Research*, vol. 83, no. 1, pp. 1–8, Feb. 2012.

[20] Y. Yang, Y. B. Liu, J. Y. Liu, Z. Huang, T. J. Liu, and G. Qiu, "Preventive transient stability control based on neural network security predictor," *Power System Technology*, vol. 42, no. 12, pp. 4076–4082, Dec. 2018.

[21] C. X. Liu, K. Sun, Z. H. Rather, Z. Chen, C. L. Bak, P. Thøgersen, and P. Lund, "A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees," *IEEE Transactions on Power Systems*, vol. 29, no. 2, pp. 717–730, Mar. 2014.

[22] Y. Xu, Z. Y. Dong, L. Guan, R. Zhang, K. P. Wong, and F. J. Luo, "Preventive dynamic security control of power systems based on pattern discovery technique," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1236–1244, Aug. 2012.

[23] Y. Z. Zhou, J. Y. Wu, L. Y. Ji, Z. H. Yu, and L. L. Hao, "Two-stage support vector machines for transient stability prediction and preventive control of power systems," *Proceedings of the CSEE*, vol. 38, no. 1, pp. 137–147, Jan. 2018.

[24] T. Q. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, 2016, pp. 785–794.

[25] D. H. Zhang, L. Y. Qian, B. J. Mao, C. Huang, B. Huang, and Y. L. Si, "A data-driven design for fault detection of wind turbines using random forests and XGBoost," *IEEE Access*, vol. 6, pp. 21020–21031, Apr. 2018.

[26] J. Y. Wang, Z. W. Sun, B. Bao, and D. Y. Shi, "Malicious synchrophasor detection based on highly imbalanced historical operational data," *CSEE Journal of Power and Energy Systems*, vol. 5, no. 1, pp. 11–20, Mar. 2019.

[27] M. H. Chen, Q. Y. Liu, S. H. Chen, Y. C. Liu, C. H. Zhang, and R. H. Liu, "XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system," *IEEE Access*, vol. 7, pp. 13149–13158, Jan. 2019.

[28] N. Li, B. L. Li, and L. Gao, "Transient stability assessment of power system based on XGBoost and factorization machine," *IEEE Access*, vol. 8, pp. 28403–28414, Jan. 2020.

[29] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 93, Aug. 2018.

[30] S. Begum, M. U. Ahmed, P. Funk, N. Xiong, and M. Folke, "Case-based reasoning systems in the health sciences: a survey of recent trends and developments," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 4, pp. 421–434, Jul. 2011.

[31] A. Adhikari, D. M. J. Tax, R. Satta, and M. Fath, "LEAFAGE: example-based and feature importance-based explanations for black-box ML models," in *Proceedings of 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019.

[32] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[33] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[35] U. G. Knight, *Power Systems Engineering and Mathematics*, New York: Pergamon, 1972.

[36] B. F. Wollenberg and W. O. Stadlin, "A real time optimizer for security dispatch," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-93, no. 5, pp. 1640–1649, Sep. 1974.

[37] F. N. Lee, J. Huang, and R. Adapa, "Multi-area unit commitment via sequential method and a DC power flow network model," *IEEE Transactions on Power Systems*, vol. 9, no. 1, pp. 279–287, Feb. 1994.

[38] A. dos Santos, P. M. Franca, and A. Said, "An optimization model for long-range transmission expansion planning," *IEEE Transactions on Power Systems*, vol. 4, no. 1, pp. 94–101, Feb. 1989.

[39] B. Stott, J. Jardim, and O. Alsac, "DC power flow revisited," *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1290–1300, Aug. 2009.

[40] T. Y. He, Z. N. Wei, G. Q. Sun, Y. H. Sun, H. X. Zang, and Q. Gao, "Modified direct current optimal power flow algorithm based on net loss equivalent load model," *Automation of Electric Power Systems*, vol. 40, no. 6, pp. 58–64, Mar. 2016.

[41] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, Boca Raton, FL, USA: Chapman & Hall/CRC, 1984.

[42] *The Standard on Power System Simulation for Security and Stability*, Q/GDW 1404–2015, 2016.

[43] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002.

[44] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[45] R. Storn and K. Price, "Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, Dec. 1997.

[46] M. A. Pai, *Energy Function Analysis for Power System Stability*, New York: Springer Science & Business Media, 1989.

Songtao Zhang received his B.S. degree in Electrical Engineering from Xi'an Jiaotong University, Xi'an, China, in 2017. He is currently pursuing his Ph.D. degree in Electrical Engineering at the Graduate School of China Electric Power Research Institute, Beijing, China. His research interests include power system analysis and AI applications in power systems.

Dongxia Zhang received her M.S. degree in Electrical Engineering from the Taiyuan University of Technology, Taiyuan, Shanxi, China, in 1992 and her Ph.D. degree in Electrical Engineering from Tsinghua University, Beijing, China, in 1999. From 1992 to 1995, she was a Lecturer with Taiyuan University of Technology. Since 1999, she has been working at China Electric Power Research Institute. She is the co-author of four books, and more than 40 articles. Her research interests include power system analysis and planning, big data and AI applications in power systems. She is an Associate Editor of the *Proceedings of the CSEE*.

Ji Qiao received his Ph.D. degree from Tsinghua University, Beijing, China, in 2018. He is currently working at the China Electric Power Research Institute, Beijing, China. His research interests include big data and AI applications in power system analysis and operations.

Xinying Wang received his Ph.D degree from Dalian University of Technology, Dalian, China, in 2015. He is a senior engineer and a member of CSEE. He is currently working as the deputy director of the Artificial Intelligence Application Research Section of China Electric Power Research Institute. His research interests primarily include applications of artificial intelligence in the field of electric power.

Zhijian Zhang received her M.S. degree from North China Electric Power University, Beijing, China, in 2001. She is currently a Senior Engineer of the State Gird Beijing Electric Power Dispatching and Control Center, Beijing, China. She is mainly devoted to power grid dispatching and control technology.